Web Mining Achieved by Question-and-Answer Websites for Managing Knowledge

Víctor Hugo Medina García[*1], Jorge Eliecer Posada Pinzón[2], José Fernando López Quintero[3]

[1,2]Faculty Enginnering, District University "Francisco José de Caldas", Bogotá D.C., Colombia

[3]Faculty Enginnering, CUN University, Bogotá D.C., Colombia

[*1]vmedina@udistrital.edu.co; [2]jorge.posada@naturasoftware.com; [3]jose_lopezq@cun.edu.co

***Abstract*-This paper presents a comparative analysis of question-and-answers websites, embedded or applied in one of the most relevant models of knowledge management with the aim to appreciate the application of web mining in order to strengthen knowledge. The applied hypothesis s is that it is possible to implement a comprehensive Organizational Knowledge Management System (KMS) based on a Q&A site if properly configured for this purpose. Finally, this paper presents a question-and-answer system architecture based on knowledge management that, in addition to visualizing the structure of its main components, describes the operation of a Q&A system.**

Keywords- Q&A Sites; Knowledge Management; Knowledge Management Systems; Web Mining

I. INTRODUCTION

Questions-and-answer (Q&A) site [1] is the formal or academic term that has conceptualizes the web 2.0 sites which focus on the exchange of questions and answers from a community such as Yahoo! Answers or Stack Overflow. Recently, there has been an increase in the amount of questions-and-answer sites [2] both as tools everyday use and as objects of study. This trend is evidenced in the thousands of questions and answers that are created daily on these sites, as well as the hundreds of scientific papers that have investigated Q&A sites in the last decade. Furthermore, knowledge management is extensively discussed; the topic is analysed in hundreds of books, and journals are even devoted exclusively to the topic. This article attempts to make an initial exploration of the possibilities of Q&A sites as Knowledge Management Systems (KMS).

II. FOUNDATIONS

A. Knowledge Management Concepts

Knowledge management is defined [3] in this context as an emerging discipline that aims to create, share and use "tacit and explicit knowledge" to support the development of individuals and communities. This has focused on the need to manage organizational knowledge and organizational learning as key mechanisms for strengthening a region, according to the vision of the future that will determine their medium- and long-term strategic development plans [4].

As pointed out by Azpiazu (2012), "society is undertaking a fundamental transformation from the industrial age into the information age. The engines of the information age are learning and knowledge" [5]. This vision suggests that a reorganization of education, as well as the redesign and redefinition of the roles and responsibilities of officers within the education system requires the reengineering of organizational processes.

This field combines new ideas with ideas that "everyone has known for a long time" [6]. In more detail, knowledge management as defined by Murray [7], referring to the business field, is a strategy that can transform the intellectual capital of a company, as the recorded information and the talents of its members, to demonstrate higher productivity, create greater value and increase the competitiveness of organizations.

B. Question-and-Answer Sites

The main entities that can be identified in a Q&A site are as follows: questions, answers, users (people who ask questions), experts (people who answer questions) and portals (software). The relations between those entities are visualized in Fig. 1.

The primary topics of recent research related to Q&A sites are as follows: question classification, question routing, answer quality, answer summarizing, user satisfaction, user/expert motivation, expert reputations, software design, and information retrieval.

DOI: 10.5963/IJCSAI0501003

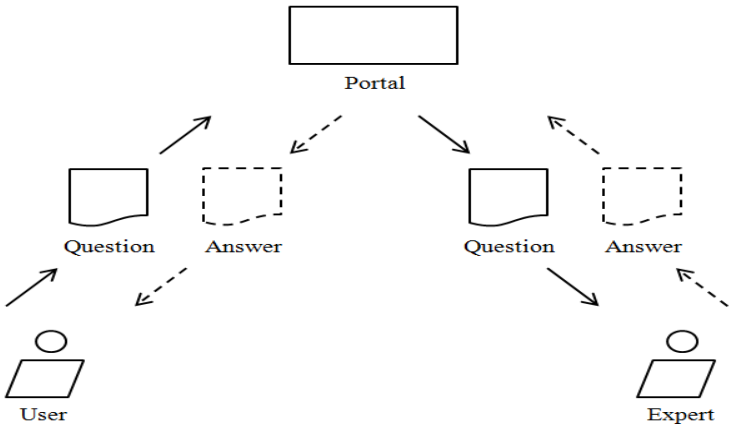

Fig. 1 Q&A site entity relations

III. DALKIR KNOWLEDGE MANAGEMENT MODEL

The knowledge management model proposed by Dalkir (2013) is perhaps the most comprehensive proposed to date regarding the so adaptation and applications of Q&A systems [8] (Fig. 2).

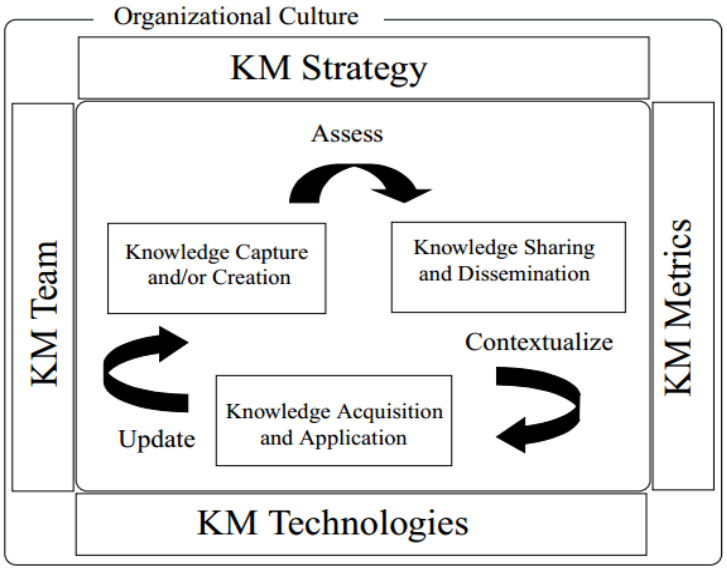

Fig. 2 Dalkir KM Model [8]

The core model is the KM Cycle, consisting of three processes: knowledge capturing and/or creation, knowledge sharing and dissemination, and knowledge acquisition and application [8]; the cycle is supported by KM Technologies. The cycle requires KM team conformation for its implementation. Continuity of the cycle must be ensured to formulate the KM strategy. To monitor and control the cycle, it is essential to define the KM metrics. All of this occurs within the context of organizational cultures.

A. Knowledge Management Cycle

The major stages of the knowledge management cycle are identified as knowledge capturing and creation, knowledge sharing and dissemination, and knowledge acquisition and application [8, 9].

1) Knowledge Capturing and/or Creation

The key phases of knowledge acquisition are conceptualization and coding [8]; these processes are interrelated, as shown in Fig. 3.

DOI: 10.5963/IJCSAI0501003

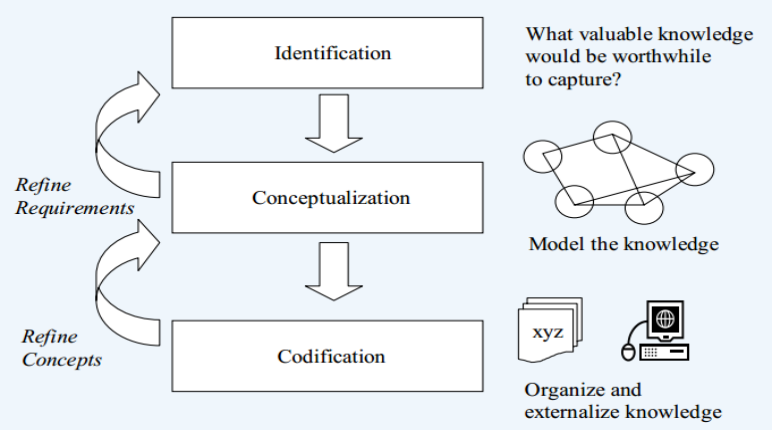

Fig. 3 Key knowledge acquisition phases [5]

The main topics of the Q&A sites related to the knowledge capturing process are question classification [10], and answer summarizing [11].

- *Question Classification*

Ignatova (2009) presented a question type classification scheme, developed to gain a better understanding of the kinds of questions people ask on social Q&A sites. They used this scheme to annotate a sample of user questions, and obtained good to very good inter annotator agreement. The annotation of opinion questions proved to be more difficult, and hence necessitates further investigation [10].

- *Answer Summarizing*

Liu, et al. (2008) determined that in the process of accumulation, cQA services assume that questions always have unique best answers. However, with an in-depth analysis of questions and answers on cQA services, the assumption cannot be true. According to analysis, at least 78% of the cQA best answers are reusable when similar questions are asked again, but no more than 48% of them are indeed the unique best answers. Analysis was conducted by by proposing taxonomies for cQA questions and answers. To better reuse the cQA content, applying automatic summarization techniques was also proposed to summarize answers. Results showed that question-type oriented summarization techniques can improve cQA answer quality significantly [11].

2) Knowledge Sharing and Dissemination

Fig. 4 depicts a generic diagram of the flow of knowledge through a Knowledge Management System (KMS): 1) An issuer user generates a knowledge request and directs it to the KMS; 2) through the KMS, the request is directed to another user; 3) the receiving user generates a knowledge response and directs it to the receiving user directly or via the KMS.

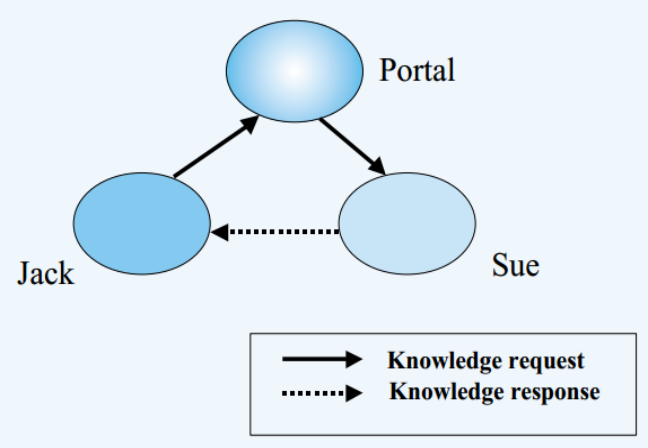

Fig. 4 Mapping the flow of knowledge [1]

If this generic KMS flow is compared to the specific flow for Q&A sites presented in Fig. 1, a one to one correspondence between the different elements can be determined, so that Q&A sites may be fully implemented using the presented abstract flow of knowledge. The primary topic of the Q&A sites related to the knowledge sharing process is question routing [12].

Zhou (2009) explained that “online forums contain huge amounts of valuable user-generated content. In current forum systems, users have to passively wait for other users to visit the forum systems and read/answer their questions” [12].

DOI: 10.5963/IJCSAI0501003

3) Knowledge Acquisition and Application

The primary topic of Q&A sites related to the knowledge application process is information retrieval [8, 13].

Jeon (2010) presented "a framework to use non-textual features to predict the quality of documents. They also show a quality measure can be successfully incorporated into the language modelling-based retrieval model" [13].

B. Knowledge Management Technologies

One difference that can be established between CMS (Content Management Systems) and KMS (Knowledge Management Systems) is granularity. In general, it can be said that in KMS, greater granularity is achieved than in CMS. Thus, KMS works well with smaller artefacts that CMS; in fact, artefacts which work best in KMS are parts or fragments of artefacts which work in CMS (documents, video, emails, etc.). This relationship is illustrated in Fig. 5. In the case of Q&A sites, artefacts consist of questions and answers that can be regarded as parts or fragments of a text.

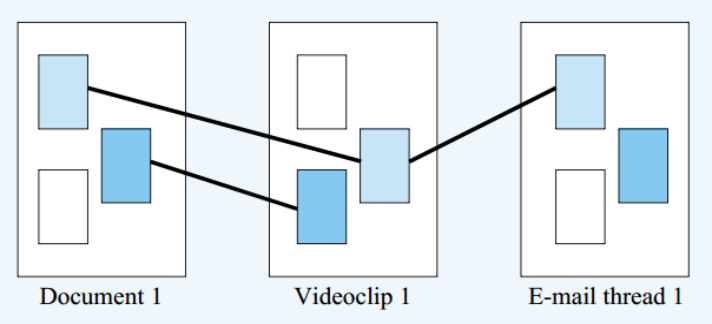

Fig. 5 Chunking in content management [8]

The primary topic of Q&A sites related to knowledge management technologies is software design [14].

Richardson and White (2011) used "data from an instant-messaging-based synchronous social Q&A service deployed to an online community of over two thousand users to study the prediction of:

- Whether a question will be answered;
- The number of candidate answerers that the question will be sent to;
- Whether the asker will be satisfied by the answer received. Predictions are made at many points of the question lifecycle (e.g., when the question is entered, when the answerer is located, halfway through the asker-answerer dialog, etc.)" [14].

C. Knowledge Management Team

The primary topic of Q&A sites related to knowledge management teams is expert reputation [15].

Jurezyk and Agichtein (2007) presented "an analysis of the link structure of a general-purpose question answering community to discover authoritative users, and promising experimental results over a dataset of more than 3 million answers from a popular community QA site" [16]. They also described "structural differences between question topics that correlate with the success of link analysis for authority discovery" [15].

D. Knowledge Management Strategy and Metrics

In his work Azpiazu (2012), states that the process of knowledge management should try to strike a balance between "Fluid" and "Institutional" trends, which are respectively related to forms of Tacit Explicit Knowledge [5]. This model is presented in Fig. 6.

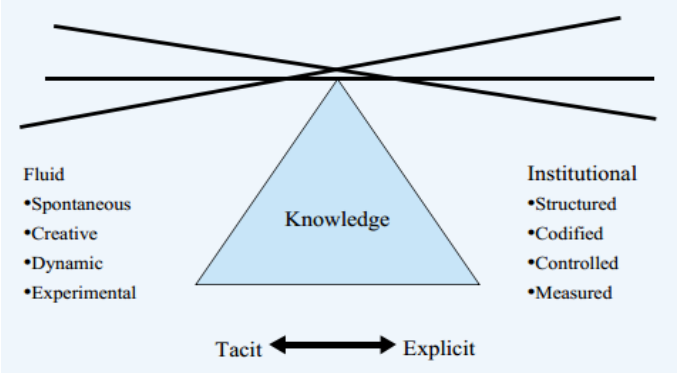

Fig. 6 Balance between fluid and institutionalization [5]

DOI: 10.5963/IJCSAI0501003

The primary topics of Q&A sites related to knowledge management strategy and metrics are answer quality and user satisfaction [11, 16, 17].

1) Answer Quality

Harper, et al. (2008) investigated "predictors of answer quality through a comparative, controlled field study of responses provided across several online Q&A sites. Along with several quantitative results concerning the effects of factors such as question topic and rhetorical strategy, they proposed two high-level messages: 1) You get what you pay for in Q&A sites; 2) A Q&A site's community of users contributes to its success" [16].

In his paper, Ackerman & McDonald (1996) "compares answer quality on four Q&A sites: Askville, WikiAnswers, Wikipedia Reference Desk, and Yahoo! Answers. Findings indicate that: 1) the use of similar collaborative processes on these sites results in a wide range of outcomes. Significant differences in answer accuracy, completeness, and verifiability were found; 2) answer multiplication does not always result in better information. Answer multiplication yields more complete and verifiable answers but does not result in higher accuracy levels; and 3) a Q&A site's popularity does not correlate with its answer quality, on all three measures" [17].

2) User Satisfaction

Liu et al (2008) formulated "a new problem of predicting the satisfaction of web searchers with CQA answers". They analyzed a large number of web searches that resulted in visits to a popular CQA site, and identified unique characteristics of searcher satisfaction in this setting, namely, the effects of query clarity, query-to-question match, and answer quality. They then proposed and evaluated several approaches to predicting searcher satisfaction that exploit these characteristics [11].

E. The Role of Organizational Culture

The primary topic of the Q&A sites related to organizational culture is users/experts motivation [18, 19].

"What are the antecedents, inhibitors and catalysts to providing information and participating in mixed fee-based and free online contexts?" Rafaeli, Raban and Ravid (2005) "describe the behaviour of participants in this system over a 29-month period. They thus corroborate some of the theories of hybrid explanation presented to date mostly in laboratory settings" [18].

Dror, Maarek and Szpektor (2013) "propose to provide a type of "heads up" to askers by predicting how many answers, if at all, they will get. Giving a pre-emptive warning to the asker at posting time should reduce the frustration effect and hopefully allow askers to rephrase their questions if needed. To the best of their knowledge, this is the first attempt to predict the actual number of answers, in addition to predicting whether the question will be answered or not. To this effect, they introduce a new prediction model, specifically tailored to hierarchically structured CQA sites" [19].

IV. Q&A SYSTEM'S ARCHITECTURE

Question-and-answer system architecture based in knowledge management aims to formalize the integration and interaction of websites with systems or new knowledge management applications. The recent evolution and trend of web technologies in the support so-called "Web Services" confirms that this architecture strengthens the integration and support of intelligent agents. Therefore, Fig. 7 conceptually schematizes its adoption, based on the state-of-the-art theme of agents, their interactions on the web and document management support to answer questions from the community.

A. Question & Answer Systems in Knowledge Management

A Q&A system receives questions and then conducts searches to find answers to achieve a query and analyze a variety of source documents, and thus extract and formulate concise answers. The system is supported or accesses different knowledge management applications according to user questions, and responds appropriately to the interaction of intelligent agents accessing a repository of knowledge (documents), as explained below.

B. Intelligent Agents Query

The agents in this component are responsible for answering questions of knowledge and power the integrated repository of knowledge from user interaction. Clearly we can expect behaviors as being able to manage these consultation agents.

Agents access the repository of integrated knowledge and search for the most relevant answer to a question from a user or user system knowledge application. The intelligent of the agents acts when each agent learns how to best respond to the users questions. They should ideally be able to process natural language.

The other role or behavior of agents is to feed the knowledge base; in this case, the agent receives explicit knowledge of the user, and their function is to classify and categorize according to the ontology of the integrated knowledge repository.

Access to agents is achieved by a web services interface and uses ACL messages to ensure interoperability with other applications, particularly those web services associated with semantic ontology reflecting the integrated knowledge repository.

DOI: 10.5963/IJCSAI0501003

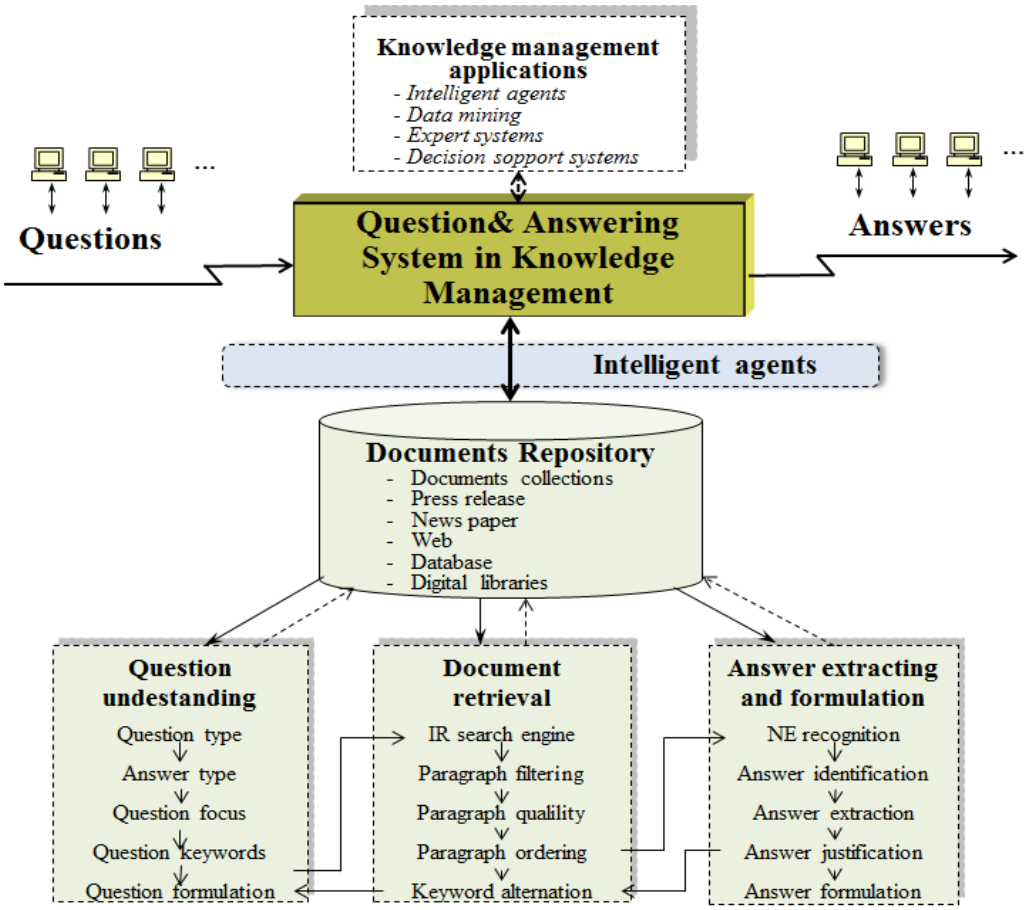

Fig. 7 Q&A system architecture based on knowledge management. Source: Authors

C. *Document Repository*

This component consists of a knowledge base that may contain structured or semi-structured (documents) data that can be generally described by any ontology.

This repository can be powered by other specific repositories using data mining techniques, or can be powered directly by users through intelligent agents acting as consultants.

This component should also include indexers, cataloguers and other unified search tools.

D. *Operation of Q&A System*

According to Moldovan (2000), a Q&A system receives questions and then conducts searches to find answers in order to achieve a query and analyze a variety of source documents, and thus extract and formulate concise answers [20].

In accordance with Fig. 7, the essential modules in most Q&A systems are as follows:

- Understanding of the question;
- Document recovery;
- Extraction and formulation of response (Fig. 5).

The compression modulus of the question "determines the type of question and the type of response expected, builds an approach to answer, and the question becomes a series of queries to the search engine. The better the system of the intended question is understood, the easier it is to extract the answer. To find the correct answer from a large collection of texts, you first need to know" [21] what to look for.

"Response rates can usually be determined from the question. To better detect the response, the system first classifies the type of question: what, why, who, how, where and when" [21]. However, the type of question is not sufficient to find answers. For example, with the question: "Who was the first American in space?" The kind of answer is obvious: people. However, this does not apply to questions based on "what", because these questions are more ambiguous. "The same applies to many other types of questions; however, this problem is solved by defining a concept approach. One approach is a word or sequence of words that define the issue and eliminates the ambiguity stating what the question is about something". For example, to the

DOI: 10.5963/IJCSAI0501003

question "What it is the largest city in Spain?" [21], the approach is the largest city.

"The document retrieval module is a search engine that extracts relevant documents from a collection of documents. After extracting the relevant documents, the module identifies paragraphs that contain possible answers" [22]. This decreases the amount of text that must be analyzed in order to extract answers.

In the extraction module and the formulation of responses, "one or more pieces of information will eventually be used to formulate the answer". This must rely on lexical-semantic information, which is provided by a scanner that identifies named entities, currencies, dates and time expressions or location, and products. Finally, "recognizing the type of response, through the semantic label returned by the (intelligent) agent, creates a candidate answer". The extraction module based on the response evaluates a set of heuristics. The best search engine on the web can limit the amount of text responses, i.e., the module performs less work. Some systems apply a justification to perform logical response profiles [22].

V. CONCLUSIONS AND FUTURE WORK

The Knowledge Management (KM) model [10] is notable for its incorporation of the elements of the most widely-used and recognized KM models that have been previously proposed. Q&A sites can be considered a full implementation of this KM model, because these sites can identify each and every one of the elements of the KM model. In ontological terms, one could argue that Q&A sites represent individuals instances of the abstract concept or abstract class of Knowledge Management Systems (KMS), and therefore these sites provide good laboratory in which to study and experiment with this concept.

Proposed future work is to perform a comparative analysis of the Wiki sites (such as Wikipedia) with the referred KM model, and then make a proposal that combines elements of Q&A sites and Wiki sites to maximize the use and exploitation of Knowledge Management Systems (KMS) in organizations.

REFERENCES

[1] Sanghee Oh, "The Characteristics and Motivations of Health Answerers for Sharing Information, Knowledge, and Experiences in Online Environments," *Journal of the American Society for Information Science and Technology*, pp. 543-557, March 2012. DOI:10.1002/asi.21676. http://www.researchgate.net/profile/Sanghee_Oh/publications.

[2] Soojung Kim and Sanghee Oh, "Users' relevance criteria for evaluating answers in a social Q & A site," *Journal of the American Society for Information Science and Technology*, pp. 716-727, April 2009. DOI:10.1002/asi.21026. http://www.researchgate.net/profile/Sanghee_Oh/publications/2.

[3] R. Young, B. Bunyagidj, S. Kim, P. Nair, N. Ogiwara, I. Yasin, and S. Talisayon, "Knowledge Management for the Public Sector," *Asian Productivity Organization (APO)*, Tokyo, 2013. http://www.apo-tokyo.org/publications/ebooks/page/2/ https://www.tanum.no/serie/Intelligent%20Systems%20Reference%20Library.

[4] V. H. Medina, J. A. Gil, and D. Liberona, "Knowledge management model as a factor of educative quality: Towards an excellence model," *Journal LNBIP 185 - Lecture Notes in Business Information Processing*, Ed. Springer-Verlag Berlin, pp. 78-89, 2014.

[5] J. Azpiazu, J. Pazos, and A. Silva, "A virtual classroom based on academic memories," In: *Proceedings of International Conference on Information and Communication Technologies in Education (ICTE2002)*, Spain, 2002.

[6] L. Prusak, *Working knowledge: How Organizations manage what They Know*, Boston: Harvard Business School Press, 2010.

[7] T. Murray, T. Shen, J. Piemonte, C. Condit, and J. Thibedeau, "Adaptivity for Conceptual and Narrative Flow in Hyperbooks: the Metalinks System," in *Proceedings of Adaptive Hypermedia 2000*, Trento Italy, August 2000. [Disp. Online: http://helios.hampshire.edu/~ tjmCCS/papers/AHM2000murray. doc].

[8] K. Dalkir, *Knowledge management in theory and practice*, Routledge, 2013.

[9] J. Braton and J. Gold, *Human Resource Management: Theory and Practice*, 2nd ed., Lawrence Erlbaum Associates, Inc., Pub. Mahwah, New Jersey, 2000. Online: www.info2myfriends.blog.com.

[10] K. Ignatova, C. Toprak, D. Bernhard, and I. Gurevych, "Annotating question types in social Q&A sites," In *Tagungsband des GSCL Symposiums 'Sprachtechnologie und eHumanities*, pp. 44-49, 2009. https://www.ukp.tu-darmstadt.de/publications/details/?no_cache=1&tx_bibtex_pi1%5Bpub_id%5D=TUD-CS-2009-0005.

[11] Y. Liu, S. Li, Y. Cao, C. Y. Lin, D. Han, and Y. Yu, "Understanding and summarizing answers in community-based question answering services," In *Proceedings of the 22nd International Conference on Computational Linguistics-Volume, Association for Computational Linguistics*, pp. 497, 2008.

[12] Y. Zhou, G. Cong, B. Cui, C. S. Jensen, and J. Yao, "Routing questions to the right users in online communities," In *Data Engineering, 2009. ICDE'09. IEEE 25th International Conference*, pp. 700-711, 2009.

[13] G. Y. Jeon, Y. M. Kim, and Y. Chen, "Re-examining price as a predictor of answer quality in an online Q&A site," In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, ACM*, pp. 325-328, 2010.

[14] M. Richardson and R. W. White, "Supporting synchronous social q&a throughout the question lifecycle," In *Proceedings of the 20th international conference on World Wide Web, ACM*, pp. 755-764, 2011.

[15] P. Jurczyk and E. Agichtein, "Discovering authorities in question answer communities by using link analysis," In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pp. 919-922, 2007.

[16] F. M. Harper, D. Raban, S. Rafaeli, and J.A. Konstan, "Predictors of answer quality in online Q&A sites," In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, ACM*, pp. 865-874, 2008.

DOI: 10.5963/IJCSAI0501003

[17] M. S. Ackerman and D. W. McDonald, "Answer Garden 2: merging organizational memory with collaborative help," In *Proceedings of the 1996 ACM conference on Computer supported cooperative work, ACM*, pp. 97-105, 1996.

[18] S. Rafaeli, D. R. Raban, and G. Ravid, "Social and economic incentives in Google Answers," In A*CM Workshop Sustaining Community: The role and design of incentive mechanisms in online systems*, Sanibel Island, FL USA, 2005.

[19] G. Dror, Y. Maarek, and I. Szpektor, "Will My Question Be Answered? Predicting "Question Answerability," in *Community Question-Answering Sites. In Machine Learning and Knowledge Discovery in Databases*, Springer Berlin Heidelberg, pp. 499-514, 2013.

[20] D. Moldovan, "Question-Answering Systems in Knowledge Management," *IEEE Intelligent Systems*, vol. 16, no. 6, pp. 90-92, Dec. 2001. http://www.hlt.utdallas.edu/~moldovan/newpapers/index.html.

[21] S. Harabagiu, D. Moldovan, and J. Picone, "Open-Domain Voice-Activated Question Answering," In *Proceedings of the 19th International Conference on Computational Linguistics (COLING-2002)*, Taipei, Taiwan, pp. 321-327, August 2002. http://www.hlt.utdallas.edu/page.php?p=publications.

[22] D. Moldovan, "The Structure and Performance of an Open-Domain Question Answering System," *Proc. 38th Ann. Meeting of the Assoc. Computational Linguistics (ACL 2000)*, Morgan Kaufman, San Francisco, 2000.

Víctor Hugo Medina García was born in Santa Rosa de Viterbo, Colombia. He received the PhD degree in Computer Engineering from the Pontificia of Salamanca University, DEA Languages, Systems and Software Engineering and Master in Computer Science at the Polytechnic University of Madrid in Spain, Specialist in Marketing from the Rosario University and Systems Engineering of the District University "Francisco José de Caldas" of Colombia.

He is currently a researcher and senior lecturer at the Faculty of Engineering at the District University in Bogotá - Colombia, where he is Extension Director of Engineering Faculty, he was Director of Engineering Doctorate. It is also professor in UNIR - International University of the Rioja and he was coordinator, associate professor and visiting professor in Computer Engineering from the Pontificia of Salamanca University campus of Madrid - Spain. His area of work and research is knowledge management and software engineering.

Prof. Medina is memberships in professional societies like: Senior member of IACSIT - International Association of Computer Science and Information Technology, member No: 80338528; LACCEI (Latin American and Caribbean Consortium of Engineering Institutions); and GICOGE (International Research Group of Information, Communication and Knowledge Management).

Jorge Eliecer Posada Pinzón was born in Bogotá, Colombia. He received the degree in Computer Engineering from the ECCI University, candidate of the Master for Engineering Industrial of the District University "Francisco José de Caldas" of Colombia.

He is currently a researcher and teacher at the Faculty of Engineering at the ECCI University. His area of work and research is software engineering.

José Fernando López was born in Bogotá, Colombia. He received the degree Master in Computer Science at the Andes University, Computer Engineering from the District University "Francisco José de Caldas", and candidate of the Doctorate for Oviedo University of Spain.

He was Vice Chancellor for Planning and Quality Assurance and teacher at the Faculty of Engineering at the ECCI University and he is currently Vice Chancellor of CUN (Corporación Unificada Nacional de Educación Superior). His area of work and research is software engineering.

DOI: 10.5963/IJCSAI0501003